DOCUMENT RESUME

ED 319 231 FL 018 498

AUTHOR Stansfield, Charles W.
TITLE An Evaluation of Simulated Oral Proficiency Interviews as Measures of Spoken Language Proficiency.
PUB DATE 90
NOTE 13p.; Paper presented at the Georgetown University Roundtable on Languages and Linguistics (Washington, DC, March, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Interviews; *Language Proficiency; *Language Tests; *Simulation; *Test Format
IDENTIFIERS ACTFL Proficiency Guidelines; Oral Proficiency Interview; *Simulated Oral Proficiency Interview

ABSTRACT

A discussion of the simulated oral proficiency interview (SOPI), a type of semi-direct speaking test that models the format of the oral proficiency interview (OPI), describes its development and research and examines its usefulness. The test used for discussion is a tape-recorded test consisting of six parts, scored by a trained rater using the American Council of the Teaching of Foreign Languages (ACTFL)/Interagency Language Roundtable (ILR) proficiency scale. A review of research on different SOPI tests in different contexts reveals a high correlation with the OPI and some practical and psychometric advantages over it. The OPI must be administered by a trained interviewer, whereas any teacher, aide, or language lab technician can administer the SOPI. The SOPI can be simultaneously administered to a group of examinees by a single administrator, whereas the OPI must be individually administered. The SOPI may be preferred for some testing purposes, such as qualification for employment, and the OPI for others such as placement or program evaluation. It is concluded that the SOPI may not be, as previously characterized, only a "second-order substitute" for the OPI. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319231

An Evaluation of Simulated Oral Proficiency Interviews as Measures of Spoken Language Proficiency

Paper presented at the

1990 Georgetown University Roundtable on Languages and Linguistics

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by

Charles W. Stansfield

Center for Applied Linguistics

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G.R. Tucker
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

In his discussion of semi-direct tests, Clark says that "semi-direct tests may be proposed as second-order substitutes for direct techniques when general proficiency measurement is at issue, but it is not operationally possible to administer a direct test" (p. 48). The major purposes of this paper are to describe the development and research that has been conducted to date on the simulated oral proficiency interview (SOPI), and to examine whether Clark's characterization of semi-direct tests should extend to the SOPI.

Description of the SOPI

The simulated oral proficiency interview is a type of semi-direct speaking test that models, as closely as is practical, the format of the oral proficiency interview (OPI). The OPI is used by US Government agencies belonging to the Interagency Language Roundtable (ILR) and the American Council for the Teaching of Foreign Languages (ACTFL) to assess general speaking proficiency

FL 018448

in a second language.

The measure I have called an SOPI (Stansfield, 1989) is a tape-recorded test consisting of six parts. It begins with simple personal background questions posed on the tape in a simulated initial encounter with a native speaker of the target language. During a brief pause, the examinee records a short answer to each question. Part one is analogous to the "warm-up" phase of the OPI. The remaining five parts are designed to elicit language that is similar to that which would be elicited during the level check and probe phases of the OPI. Parts two, three, and four employ pictures in a test booklet to check for the examinee's ability to perform the various functions that characterize the Intermediate and Advanced levels of the ACTFL proficiency guidelines, or levels one and two of the ILR skill level descriptions. Thus, the examinee is asked to give directions to someone using a map, to describe a particular place based on a drawing, and to narrate a sequence of events in the present, past, and future using drawings in the test booklet as a guide. Parts five and six of the SOPI require the examinee to tailor his or her discourse strategies to selected topics and real-life situations. These parts assess the examinee's ability to handle the functions and content that characterize the Advanced and Superior levels of the ACTFL guidelines, or levels two through four of the ILR skill level descriptions. Like the OPI, the SOPI can end with a wind-down.

After the test is completed the tape is scored by a trained rater using the ACTFL/ILR scale. Scores may range from the Novice

level to High Superior. The latter score is equivalent to a rating of between 3+ and 5 on the ILR scale.

Description of semi-direct tests

As indicated above, the SOPI is a type of semi-direct test. Clark (1979) defined a semi-direct test as one that elicits speech by means of tape recordings, printed test booklets, or other non-human elicitation procedures. A semi-direct test can employ a wide variety of items formats. These may include techniques such as spoken pattern practice in response to cues in the test booklet or on tape, reading aloud, sentence repetition, sentence completion, naming nouns or verbs depicted through line drawings in the test booklet, describing a single picture or describing a picture sequence (Clark, 1979; Clark & Swinton, 1979). Many of these elicitation techniques are inherently different from the relatively authentic, context-based techniques that would be found in the OPI and in the SOPI.

Research and development involving the SOPI

In five studies involving different test development teams and different languages, the SOPI has shown itself to be a valid and reliable surrogate of the OPI. Clark and Li (1986) developed the first SOPI, although they did not label it as such, in an effort to incorporate modifications that Clark felt could improve the Recorded Oral Proficiency Interview, or ROPE test (Lowe and Clifford, 1980). Clark and Li developed four forms of a ROPE-like

test of Chinese, with instructions and scenarios in English, and then administered the four forms and an OPI to 32 students of Chinese at two universities. Each test was scored by two raters and the scores on the two types of test were statistically compared. The results showed the correlation between the SOPI and the OPI to be .93.

Shortly after arriving at the Center for Applied Linguistics (CAL) in 1986, I read Clark's report on this project and realized that these favorable results merited replication by other researchers in situations involving other test developers and learners of other languages. As a result, I applied to the International Research and Studies Program for a grant to develop similar tests in four other languages. Fortunately, the grant was funded, and in August 1987 I began the development of a similar semi-direct interview test of Portuguese, called the Portuguese Speaking Test (Stansfield, et al., 1990). Three forms of this test and an OPI were administered to 30 adult learners of Portuguese at four institutions. Each test was also scored by two raters. In this study a correlation of .93 between the two types of test was also found. In addition, the SOPI showed itself to be slightly more reliable than the OPI and some raters commented that the SOPI seemed easier to rate, since the format of the test did not vary with each examinee.

One of the things we learned as a result of our experience with the PST, was the realization that it would be possible to include a wind-down after Part VI of the test. This is usually an

easy question designed to put the examinee at ease and to facilitate the ending of the examination in as natural manner as possible (Stansfield and Kenyon, 1988). We incorporated a wind-down with the Hausa test we developed subsequently, and we plan to incorporate a wind-down in any future forms of the PST that we develop. Another thing we learned is that the SOPI may differ somewhat for each language, in order to accommodate the unique characteristics of that language. For instance, for the PST, it was necessary to record two versions of the test, one in Lusitanian Portuguese and one in Brazilian Portuguese, since in Part I each dialect proved to be quite problematic for learners who had been exposed to only one dialect, which is often the case with Portuguese instruction in the U.S.

During 1988 and 1989, I directed the development of tests in Hebrew, Hausa, and Indonesian. The Hebrew SOPI, or Hebrew Speaking Test (HeST) as we call it, was developed in close collaboration with Elana Shohamy and her associates at the University of Tel Aviv (Shohamy et al., 1989). In order to accommodate the different settings where the language is studied and used, two forms of the test were developed for use in Hebrew language schools for immigrants to Israel, and two forms were developed for use in North America. Because the pronoun "you" carries gender in Hebrew, alternate versions of the master tape for men and women were developed. The first two forms were administered to 20 foreign students at the University of Tel Aviv and the other two forms were administered to 10 students at Brandeis University and 10 students

at the University of Massachusetts at Amherst. Each group also received an OPI. The correlation between the OPI and this SOPI for the Israeli version was .90, while the correlation for the U.S. version was .94. Parallel-form and interrater reliability were also very high. The average interrater reliability was .94 and parallel form reliability was .95. When examinees' responses on different forms were scored by different raters, the reliability was .92.

Recently, Dorry Kenyon (my associate at CAL) and I reported on the development and validation of SOPIs in Indonesian and Hausa (Stansfield and Kenyon, 1989). The development of the Indonesian Speaking Test (IST) posed special problems. Indonesian is one of those languages where the context of the speech situation seems to be especially important. Because of this, we strived to contextualize the test items to an even greater degree than we had done for other languages. In order to do this, we specified the age, sex, and position or relationship of the supposed interlocutor for the examinee. During trialing, we noticed that examinees tended to assign a name to the person they were speaking with. As a result, when appropriate, we gave each interlocutor a name on the operational forms. To validate the test, 16 adult learners of Indonesian were administered two forms of the IST and an OPI. The correlation with the OPI was .94. Reliability was also high, with interrater reliability averaging .98, and parallel-form reliability averaging .94 for the two raters. When different forms and different raters were used, the reliability was also .93.

The development of two forms of the Hausa Speaking Test also posed special problems. First, it was necessary to develop a male and a female version of each master tape. In addition, because no ACTFL or ILR-certified interviewer/raters were available for Hausa, it was not possible to administer an OPI to the 13 subjects who took the Hausa Speaking Test. However, two speakers of Hausa as a second language, who had received familiarization training in English with the ACTFL/ILR scale, subsequently scored the Hausa test tapes on that scale. Although, as might be expected, the reliability of these raters was not as high as that which was obtained on the other SOPI tests using certified raters, the reliabilities were still quite good. The raters showed high interrater reliability, averaging .91 for the two forms of the test, and an average parallel-form reliability of .81. When different forms and raters were used, the correlation between scores was .84. These reliabilities are based on product moment correlations, which were derived by converting ACTFL/ILR scores to a numerical value. When the rank order correlation was employed to determine reliability, as is generally done with tests that employ an ordinal scale, the average interrater reliability was .95, parallel form reliability was .93, and parallel-form reliability using different raters was also .93. In addition, the raters indicated that they believed the Hausa SOPI elicited an adequate sample of language with which to assign a rating.

The SOPI versus the OPI

In comparison with the OPI, the SOPI would seem to offer certain advantages. The OPI must be administered by a trained interviewer, whereas any teacher, aide, or language lab technician can administer the SOPI. This may be especially useful in locations where a trained interviewer is not available. The SOPI can be simultaneously administered to a group of examinees by a single administrator, whereas the OPI must be individually administered. Thus, the SOPI may be preferable when many examinees need to be tested within a short span of time.

In addition to these practical advantages, the SOPI may offer psychometric advantages in terms of validity and reliability. The OPI typically takes 20 to 25 minutes to administer and produces 12-15 minutes of examinee speech. The SOPI takes 45 minutes to administer and produces a longer sample, usually 20-23 minutes of examinee speech. The more extensive sample may contribute to a more valid assessment.

In an OPI, the validity of the test sample elicited is in large part determined by the skill of the interviewer. Interviewers can vary considerably in their interviewing techniques, yet the SOPI offers the same quality of interview to each examinee.

The OPI also helps ensure high reliability. By recording the test for later scoring, it is possible to ensure that examinees will be rated by the most reliable raters. In the OPI, the same interviewer typically rates and scores the test. Yet this interviewer may not be the most reliable or accurate rater. Also,

some raters who have scored both types of test have reported that it is sometimes easier to assign a rating to an SOPI performance. In part, this may be because the SOPI produces a longer speech sample and because each examinee is given the same questions. Thus, it may be easier for the rater to apply the scale to a single test, as is the case with the SOPPI, than to many different tests, at the same time, as is the case with the OPI.

Conclusion

An examination of the SOPI research, which has been carried out on different subjects, and on tests of different languages produced by different test development teams, shows that the SOPI correlates so highly with the OPI that it seems safe to say that both test the same abilities. The SOPI has also shown itself to be at least as reliable as the OPI, and in some cases more so. Thus, it seems safe to conclude that it is as good as an OPI in many situations. A comparison of the advantages of each suggests that the SOPI can offer certain practical and psychometric advantages over the OPI. Thus, it may be useful to consider the circumstances that should motivate the selection of one format or the other.

Since the tasks on the SOPI are ones that can only be effectively handled by responding in sentences and connected discourse, the SOPI is not appropriate for learners below the level of Intermediate Low. Similarly, the semi-direct format of the test does not permit the extensive probing that may be necessary to

distinguish between the highest levels of proficiency on the ILR scale, such as levels 4, 4+, and 5.

The purpose of testing may also play a role in the selection. If the test is to have very important consequences, it may be preferable to administer an SOPI, since it provides control over reliability and validity of the score. Such a situation might be found in the use of a proficiency score to determine whether or not applicants are qualified for employment, such as for teacher certification purposes. I should mention that the Texas Education Agency agrees with me on this point, since it recently decided to award CAL a contract to develop SOPI tests in Spanish and French for teacher certification purposes in Texas. On the other hand, if scores are to be used for placement within an instructional program and a competent interviewer is available, it would seem preferable to administer an OPI. In such a situation, an error in placement can be easily corrected. Similarly, an OPI administered by a competent interviewer would seem preferable for program evaluation purposes because of the qualitative information it can provide and because the score will not have important repercussions on the examinee.

Given all of the above advantages that accrue to the SOPI, it seems time to reconsider Clark's characterization of semi-direct tests as "second order substitutes" for the direct OPI. While this characterization may be applicable to semi-direct tests in general, it does not seem to apply to the SOPI.

References

Clark, J.L.D. (1979). Direct vs. semi-direct tests of speaking ability. In E.J. Briere & F.B. Hinofotis (Eds.), Concepts in language testing: Some recent studies (pp. 35-49). Washington, DC: Teachers of English to Speakers of Other Languages.

Clark, J.L.D. & Li, Y. (1986). Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 278 264).

Clark, J.L.D. & Swinton, S.S. (1979). Exploration of speaking proficiency measures in the TOEFL context (TOEFL Research Report 4). Princeton, NJ: Educational Testing Service.

Lowe, P. & Clifford, R.T. (1980). Developing an indirect measure of overall oral proficiency. In, J.R. Frith, Editor, Measuring spoken language proficiency. Washington, DC: Georgetown University Press.

Shohamy, E., Gordon, C., Kenyon, D.M., & Stansfield, C.W. (1989. The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. Bulletin of Hebrew Higher Education, 4(1), pp. 4-9.

Stansfield, C.W. (1989). Simulated oral proficiency interviews. ERIC Digest. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

Stansfield, C.W. & Kenyon, D.M. (1988). Development of the Portuguese Speaking Test. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 296 586).

Stansfield, C.W. & Kenyon, D.M. (1989). Development of the Hausa, Hebrew, and Indonesian Speaking Tests. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service, Forthcoming).

Stansfield, C.W., Kenyon, D.M., Paiva, R., Doyle, F., Ulsh, I., & Cowles, M.A. (1990). The development and validation of the Portuguese Speaking Test. Hispania, 73(3).

Table 1. Concurrent validity product moment correlations between the SOPI and the OPI.

	Same Rater	Separate Raters	Average
Chinese	.96	.90	.93
Portuguese	.93	.93	.93
Hebrew (USA)	.94	.94	.94
Hebrew (Israel)	.90	.90	.90
Indonesian	.95	.94	.94
Hausa	n/a	n/a	

Table 2. Interrater agreement (product moment correlations) in six SOPI studies.

	Within Forms (interrater reliability	Across Forms (parallel form reliability	Different forms and raters
Chinese	.92	.96	.91
Portuguese	.96	.97	.96
Hebrew (USA)	.93	.96	.92
Hebrew (Israel)	.95	.94	.93
Indonesian	.98	.94	.93
Hausa	.91	.81	.84